LEARNING THROUGH "CHEATING" WITH GAI

IT-VEST SIG WEBINAR

SCHOOL OF COMMUNICATION AND CULTURE
AARHUS UNIVERSITY

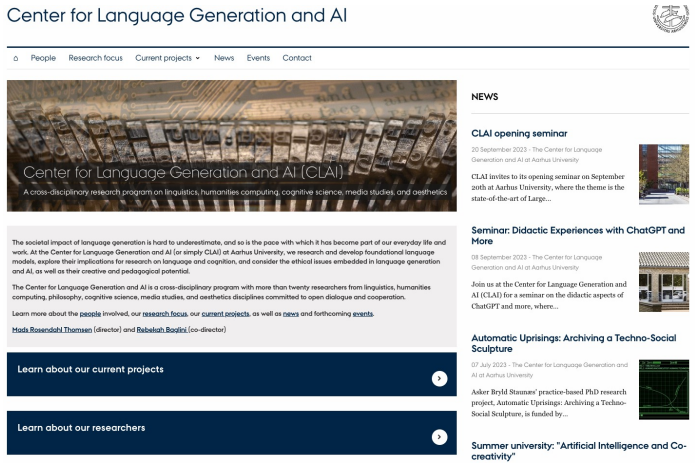
REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR

ABOUT ME

- 🇺🇸 Immigrant in 🇩🇰
- **Postdoc:** Stanford University
- **Phd, MA:** University of Chicago

- Associate professor at **AU Linguistics, Cognitive Science, & Semiotics**
  - Natural Language Processing
  - Computational Linguistics
  - Computationally-assisted media analytics
  - Semantics and pragmatics
- Founding Co-director of the **Center for Language Generation and AI (CLAI)**
- Affiliated researcher with **Center for Humanities Computing**

rbkh.net
@rebekahbaglini



ORIGINAL WORK?

A new grey area

PEN & PAPER

Exam paper
In this paper I will investigate ...

SPELL-CHECKING

Exam paper
In this papre I will investigate ...

ASSISTIVE WRITING

Exam paper
In this paper I will investigate

PARAPHRASING

Prompt: Can you help me make this paragraph more concise?

FULL GENERATION

Prompt: Please write my exam paper for me which should be about ...

PLAGIARISM

~~Anders Andersen~~
Kasper Fyhn
Exam paper
In this paper I will investigate ...

Self ← **Agency** → Other

GAI AND STUDENTS

- GAI tools are here to stay and accessible to our students
- Most exam formats cannot prevent students' use

Options

- **Ban GAI** and adopt adversarial surveillance tools to detect illicit use
- **Permit GAI** (with some constraints) and include students in conversations about academic integrity with GAI

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR

ADVERSARIAL 'AI DETECTION' SOFTWARE

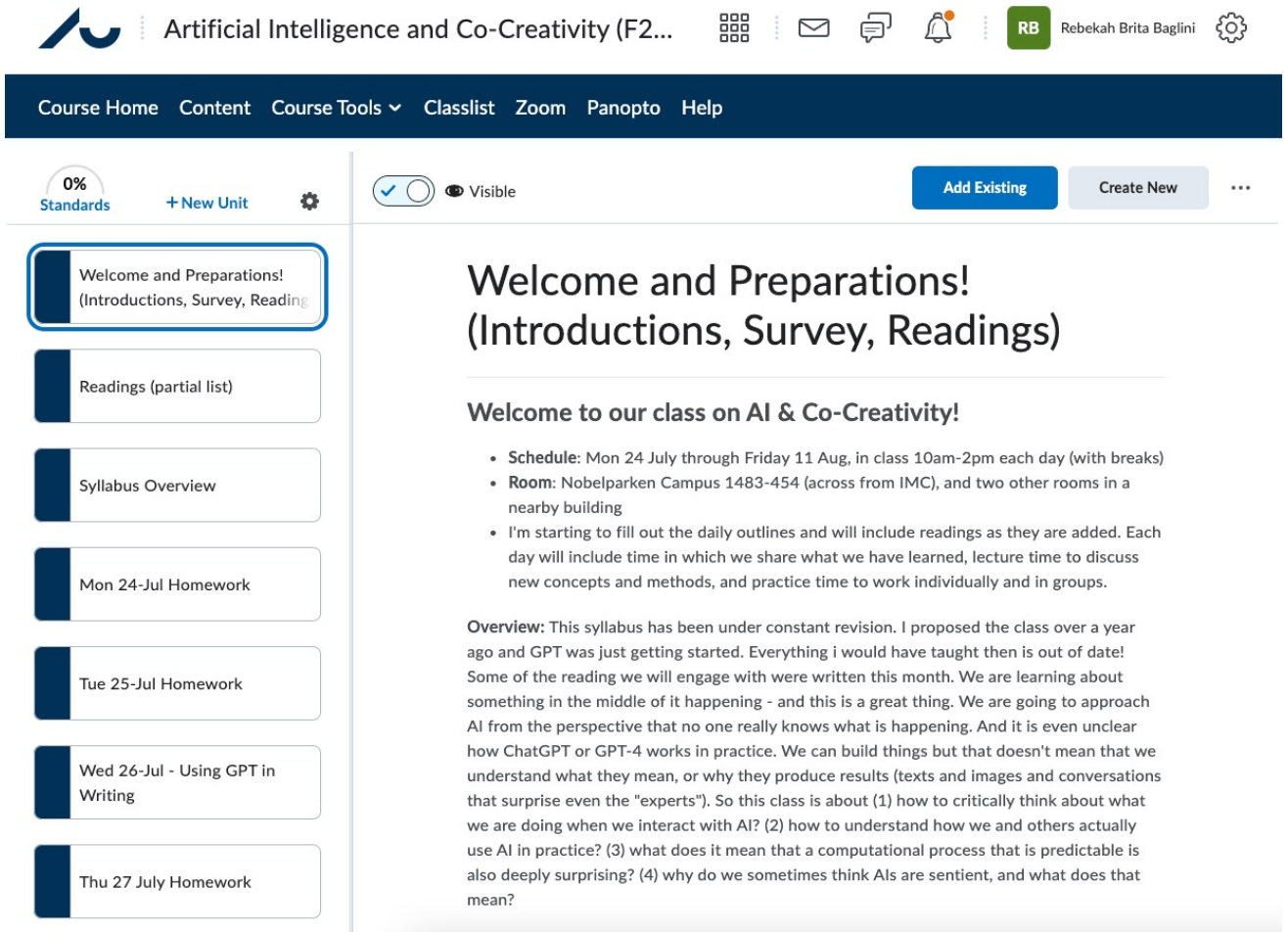
AS A CLASSROOM ACTIVITY

SCHOOL OF COMMUNICATION AND CULTURE
AARHUS UNIVERSITY

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR

AI AND CO-CREATIVITY (AU, SUMMER 2023)

- 30 1y ARTS students
- 5 instructors
  - Andreas Roepstorff
  - Joe Dumit
  - Rebekah Baglini
  - Arthur Hjorth
  - Kat Heimann



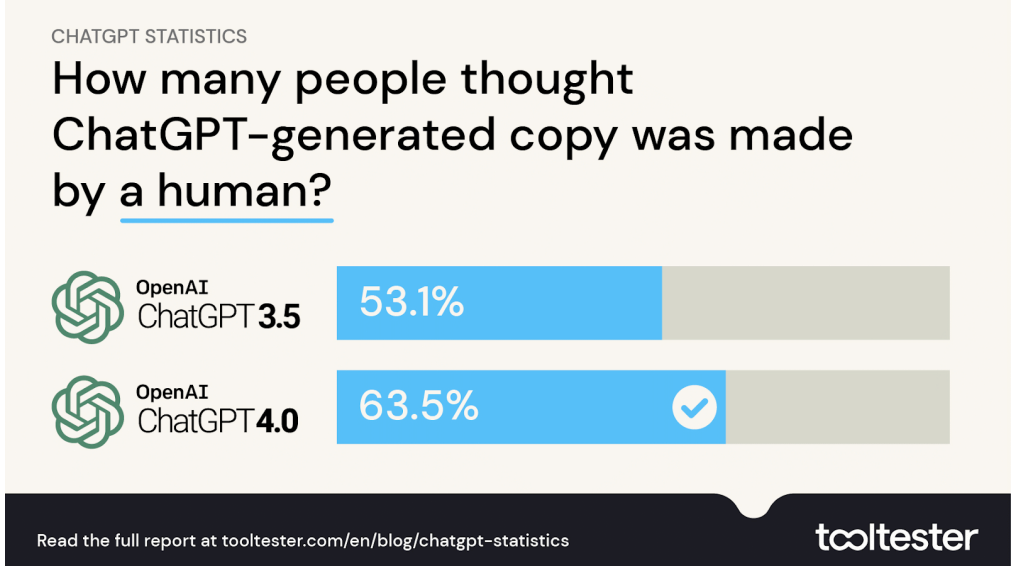
QUESTION

Do you think you can tell the difference between a text written a by a **human** and a text generated by an **AI?**

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR

QUESTION

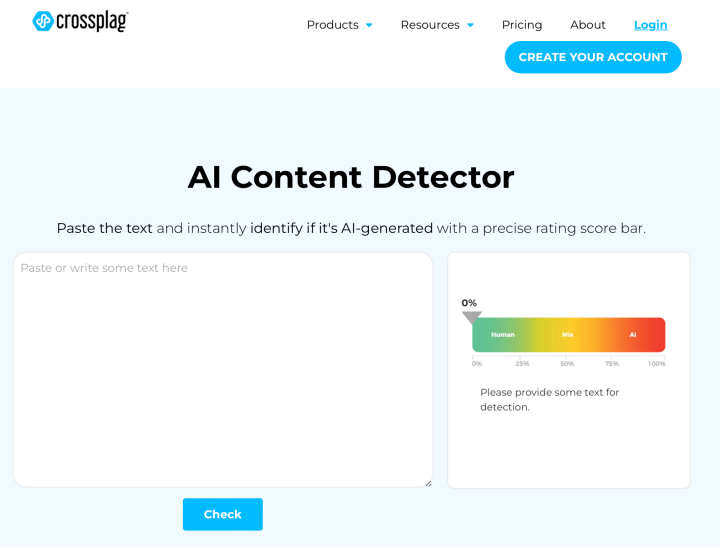
Do you think you can tell the difference between a text written a by a **human** and a text generated by an **AI?**

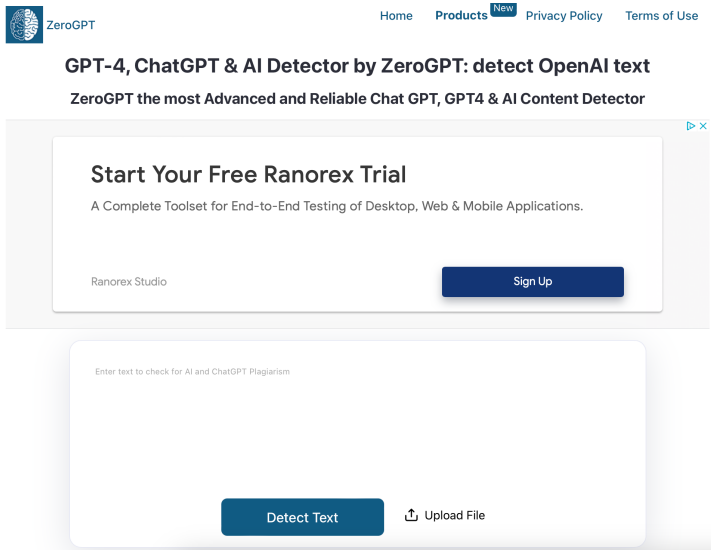


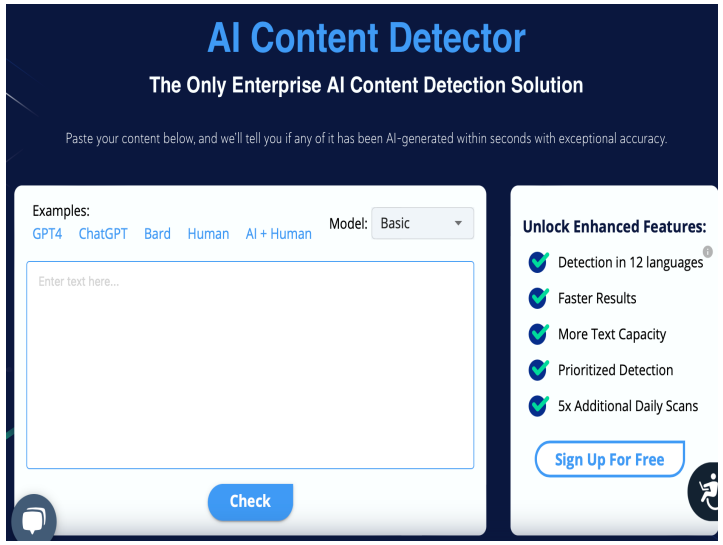
REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR

ALREADY A BOOMING BUSINESS

Some commercial software already promises to do this, e.g. **ZeroGPT, CrossPlag...**







REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR

RELIABLE?

"As of July 20, 2023, the AI classifier is no longer available due to its low rate of accuracy. We are working to incorporate feedback and are currently researching more effective provenance techniques for text, and have made a commitment to develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated."

OpenAI

As of July 20, 2023, the AI classifier is no longer available due to its low rate of accuracy. We are working to incorporate feedback and are currently researching more effective provenance techniques for text, and have made a commitment to develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated.

We've trained a classifier to distinguish between text written by a human and text written by AIs from a variety of providers. While it is impossible to reliably detect all AI-written text, we believe good classifiers can inform mitigations for false claims that AI-generated text was written by a human: for example, running automated misinformation campaigns, using AI tools for academic dishonesty, and positioning an AI chatbot as a human.

Our classifier is not fully reliable. In our evaluations on a "challenge set" of English texts, our classifier correctly identifies 26% of AI-written text (true positives) as "likely AI-written," while incorrectly labeling human-written text as AI-written 9% of the time (false positives). Our classifier's reliability typically improves as the length of the input text increases. Compared to our previously released classifier, this new classifier is significantly more reliable on text from more recent AI systems.

We're making this classifier publicly available to get feedback on whether imperfect tools like this one are useful. Our work on the detection of AI-generated text will continue, and we hope to share improved methods in the future.

Try our free work-in-progress classifier yourself:

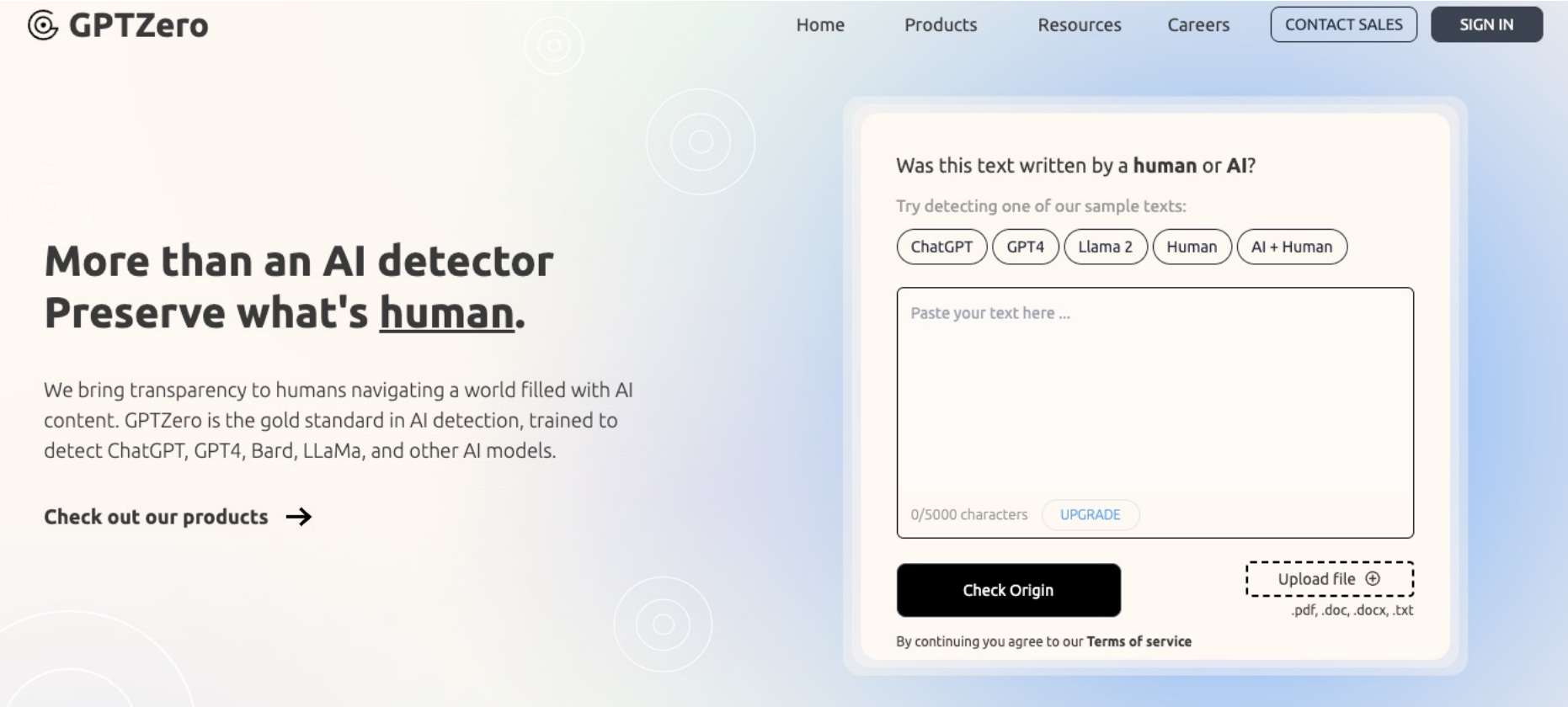
Try the classifier ↗

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR

GPTZERO.ME

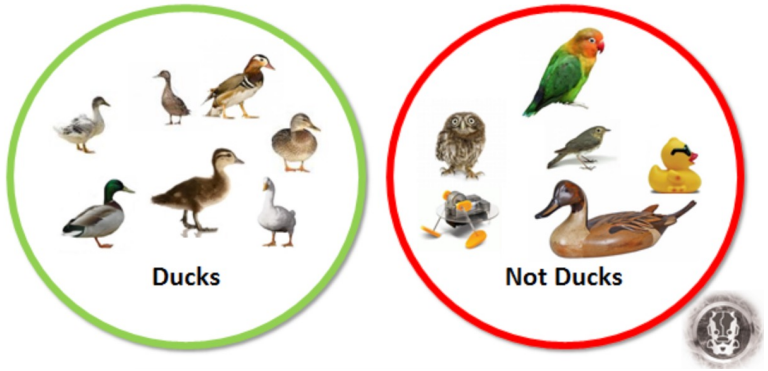
What's a classifier?

Here I focus on industry leader ZeroGPT.me and their commercial **classifier**, already used by many higher education institutions



CLASSIFICATION 101

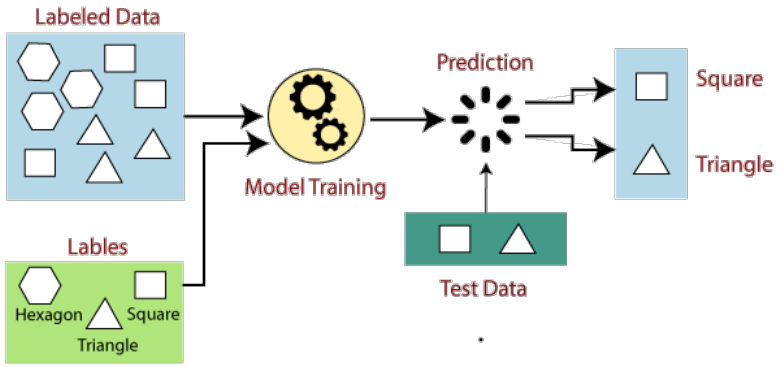
Features



Few- and zero-shot learning

"Language models are incredible few-shot learners, increasingly allowing us to bypass the laborious supervised learning route"

Supervised learning




```
Extract semantic triples from the following tweets and put them in the form (Subject)(VerbPhrase)(Arg):
| Tweet | Subject | Predicate | Object |
| --- | --- | --- | --- |
| @troelsjohnsen: @MZaccarin @PiaOlsen @SFpolitik Jeg tror vi er ret enige - og som jeg læser den her artikel, så er ham eksperten også enig.
De nye restriktioner vil have en begrænset effekt, da de eksisterende restriktioner allerede effektivt har begrænset social aktivitet, og det
er nu op til os alle at begrænse smitten. | Jeg | tror | vi er ret enige |
| @troelsjohnsen: @MZaccarin @PiaOlsen @SFpolitik Jeg tror vi er ret enige - og som jeg læser den her artikel, så er ham eksperten også enig.
De nye restriktioner vil have en begrænset effekt, da de eksisterende restriktioner allerede effektivt har begrænset social aktivitet, og det
er nu op til os alle at begrænse smitten. | vi | er | ret enige |
| @troelsjohnsen: @MZaccarin @PiaOlsen @SFpolitik Jeg tror vi er ret enige - og som jeg læser den her artikel, så er ham eksperten også enig.
De nye restriktioner vil have en begrænset effekt, da de eksisterende restriktioner allerede effektivt har begrænset social aktivitet, og det
er nu op til os alle at begrænse smitten. | som jeg | læser | den her artikel |
| @troelsjohnsen: @MZaccarin @PiaOlsen @SFpolitik Jeg tror vi er ret enige - og som jeg læser den her artikel, så er ham eksperten også enig.
De nye restriktioner vil have en begrænset effekt, da de eksisterende restriktioner allerede effektivt har begrænset social aktivitet, og det
er nu op til os alle at begrænse smitten. | eksperten | er | også enig |
```

ACTIVITY: FOOLING AI DETECTION SOFTWARE

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR

WHAT CLUES DO AI DETECTORS USE?

Perplexity measures how complex or “surprising” a sentence is. Lower the perplexity = more predictable. LMs tend to produce lower perplexity scores at the word and sentence level than humans.

Burstiness compares the variation between sentences. LMs tend to show more consistency than humans in, e.g., sentence length, complexity.

The presupposition of AI detection software:

The lower the values for these two factors, the more likely it is that a text was produced by an AI.

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR

MISSION: FOOL AI DETECTION SOFTWARE

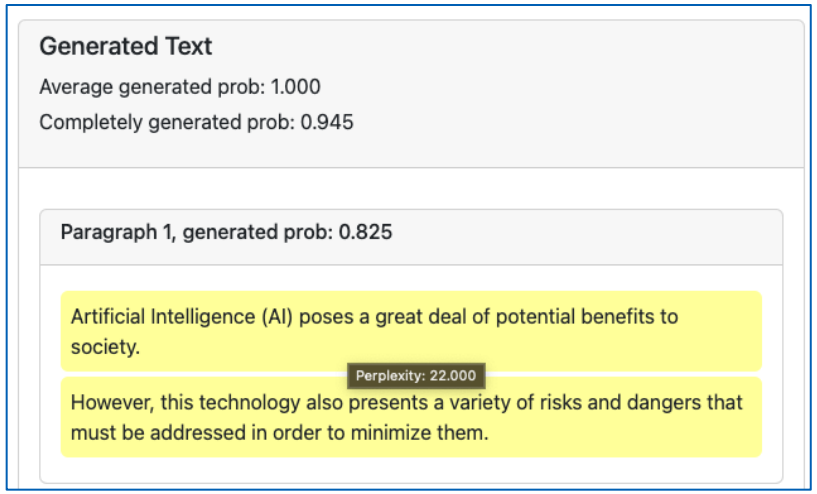
1. Go to Vortext
2. Prompt ChatGPT to write a short essay on "dangers of AI"*.
3. Check out the scores on sents and doc provided by GPTZero.
4. Make adjustments to your original prompt to try to derive lower (i.e. less likely-to-be-by-an-AI) scores
5. Reflect on your prompting: what works and what doesn't (what does/doesn't bring the score down), experiment, iterate
6. Who can get the lowest score?!

*Based on your readings for this week, you might already be familiar with this topic and have some sense a good essay on this should include.

There is a PRIZE!

INTERFACING WITH ZEROGPT API

Mouseover sentences to get perplexity scores



Sentence statistics

sentences array[object]

Information about each sentence is contained in this array, and the sentences in the document are listed in order.

- **sentence** string
- **perplexity** number

  The lower the perplexity, the more likely an AI would have generated this sentence
- **generated_prob** number

  The probability that this sentence was generated by an AI. Our current model predicts 0/1 labels, but this may change to be a percentage in the future.

Document statistics

average_generated_prob number

The average of the probabilties that each sentence was generated by an AI

completely_generated_prob number

The probability that the entire document was generated by an AI

overall_burstiness number

The amount of variation in the perplexity of the document. A useful indicator to distinguish AI and human written text

USING VORTEXT

Prompt to GPT

Input Text

Generated Text

Average generated prob: 1.000

Completely generated prob: 0.945

Paragraph 1, generated prob: 0.825

Artificial Intelligence (AI) poses a great deal of potential benefits to society.

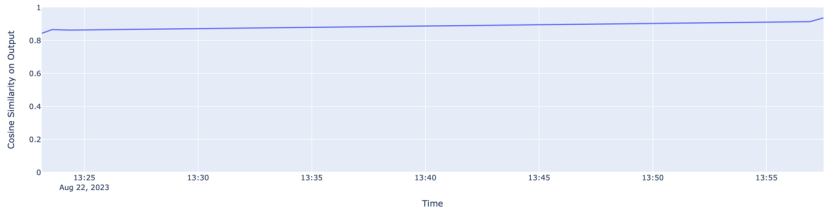
Perplexity: 22.000

However, this technology also presents a variety of risks and dangers that must be addressed in order to minimize them.

GPT output

Generate Text

Reflection



REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR

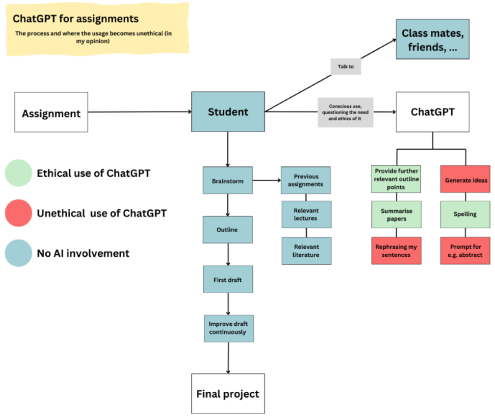
DISCUSS: YOUR PROMPTS

- What did you **change in the prompt** to get your lowest score?
- **How** did you come up with that?
- How did it seem to impact **perplexity and burstiness**?
- Did any of these changes **impact accuracy or promote hallucination?**
- Did your strategy for getting a low score sacrifice **plausibility**?

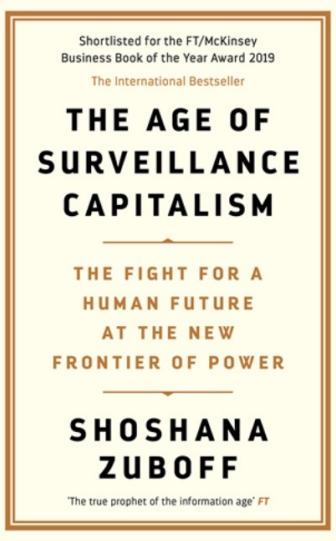
A PRIZE WAS AWARDED

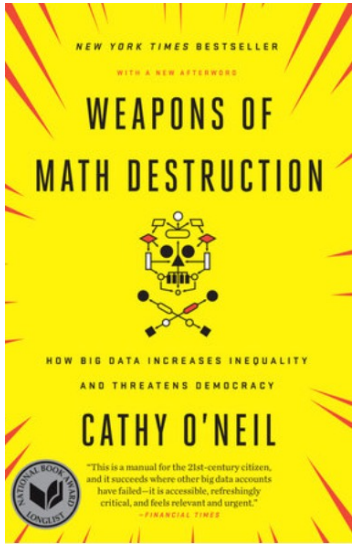
DISCUSS: BIGGER PICTURE

Activity: Diagram ethical/unethical uses of ChatGPT for assigments
Slides on students' diagrams



- How is using generative AIs as a writing tool like or unlike **traditional plagiarism**?
- Are there **ways of using ChatGPT** that are more plagiarism-like and other ways that are more okay?
- Would you be comfortable with the university adopting **mandatory AI-detection scoring**?





OUTCOMES

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR

PAPER IN PREP

"Experimenting with (il)legitimate AI Use in Universities: A Transformative Approach from Detection to Dialogue"

(Baglini, Dumit, Roepstorff, Hjorth, with 30+ student co-signers)

OUR CONCLUSIONS AND RECOMMENDATIONS

Current university policies inadequately prepare students to use GAI effectively while maintaining academic integrity.

- **Learning to work with LLMs is not restricted to programmers**
  - Prompting is a learnable skill, can build "computational thinking" as well as LLM literacy
- **AI policies cannot be one-size-fits-all**
  - What is legitimate and illegitimate will vary across disciplines and contexts
- **AI cheat detection software is worse than doing nothing**
  - It does not work, can be easily cheated, and makes students into adversaries
- **Invite students into realizing the learning objectives – with and without AI tools**
  - Afford them agency in their educational practice and learning, rather than creating an adversarial relationship between instructors and students based on mistrust

AARHUS
UNIVERSITY

THE FUTURE OF DETECTING AI-GENERATED TEXTS

SCHOOL OF COMMUNICATION AND CULTURE
AARHUS UNIVERSITY

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR

ISSUES WITH CURRENT AI-TEXT DETECTION

- Unsatisfactory performance
- Often model-specific (e.g. just for ChatGPT)
- Often blackboxed/closed source
- Uninterpretable to human users

This makes them unusable for any real-world applications that require precision and accountability, such as flagging AI-generated essays

SIDENOTE: “ECHO” PROJECT

A Scalable and Explainable Approach to Discriminating Between Human and Artificially Generated Text

- Is there more to it than **perplexity** and **burstiness**?
- If so, which **linguistic and cognitive properties** characterize artificially generated text?
- Can these features be used to build **model-independent, explainable algorithms** that reliably discriminate between human and artificially-generated text?

Roberta Rocca
Assistant Professor
roberta.rocca@cas.au.dk
1483, 422
+4587169126
+4593920366

Rebekah Brita Baglini
Associate Professor
rbkh@cc.au.dk
1483, 323

Yuri Bizzoni
Postdoc
yuri.bizzoni@cc.au.dk
1580, 236

Ross Deans Kristensen-McLachlan
Assistant Professor
rdkm@cas.au.dk

AARHUS
UNIVERSITY